Homework 4

Mark Schulist

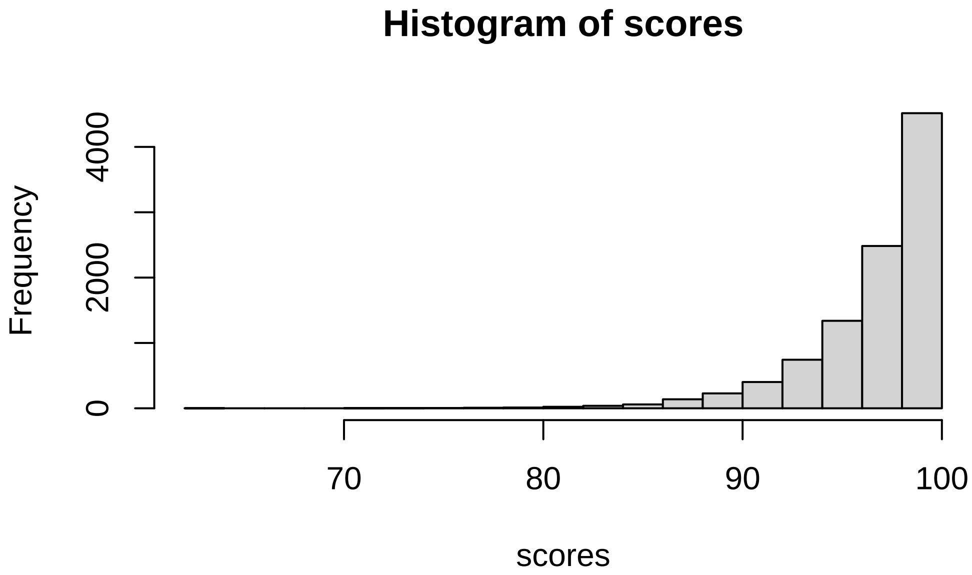
1.

a. Proportion, either they worry or don't worry.
b. Mean, revenue is continuous.
c. Proportion, they either use geolocation services or they don't.
d. Proportion, they either use web-based taxis or they don't.
e. Mean, the number of times they used web-based taxis is continuous.

2.

```r
set.seed(4)
scores <- 100 - rexp(10000, 0.3) # scores is the population

mean <- mean(scores)
sd <- sd(scores)
hist(scores)
```



The population mean is 96.644 and the population standard deviation is 3.392.

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

```
SE <- sd / sqrt(100)
```

The standard error is 0.339. Because we are taking a lot of samples from a symmetric distribution, we can assume that the CLT will hold. We want to find the probability that the sample mean ($\bar{x}$) will be within 0.4 of the population mean.

```
between <- pnorm((mean + 0.4), mean, SE) - pnorm((mean - 0.4), mean, SE)
between
```

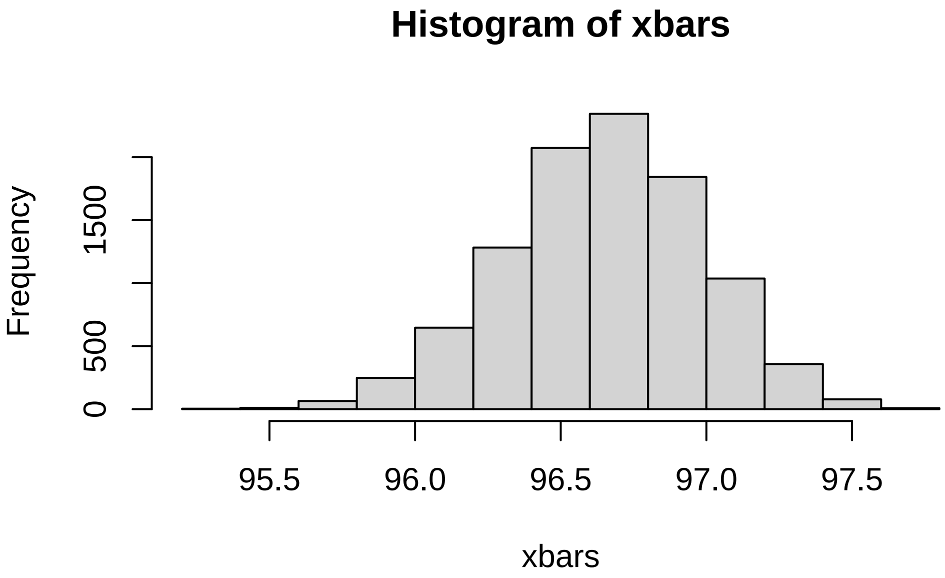
```
[1] 0.7616446
```

The probability that the sample mean is within 0.4 of the population mean is 0.762.

```
n <- 100
samples <- 10000
xbars <- rep(0, samples)

for (i in 1:samples) {
  xbars[i] <- mean(sample(scores, n))
}

m <- mean(xbars)
s <- sd(xbars)
hist(xbars)
```



Using our simulation, we found that the mean of the samples is 96.645 and the standard deviation is 0.337.

This agrees with our standard error calculation, which shows that the standard error $(SE_{\bar{x}})$ is 0.339. 0.337 is very close to 0.339. The means are also nearly identical.

We can compute the proportion of xbars that are within 0.4 of the population mean to confirm our results.

```
prop <- mean((xbars > (mean - 0.4)) & (xbars < (mean + 0.4)))
prop
```

```
[1] 0.7666
```

This agrees with our previous results that showed that the proportion of xbars within 0.4 of the population mean is 0.762.

3.

a. The Central Limit Theorem will hold with $np \geq 10$ and $n(1-p) \geq 10$. In our case, because $p = 0.05$, we would need n to be at least 200, as $200 \cdot 0.05 = 10$.
b.

```
p = 0.05
n = 500
SE = sqrt((p * (1-p)) / n)
```

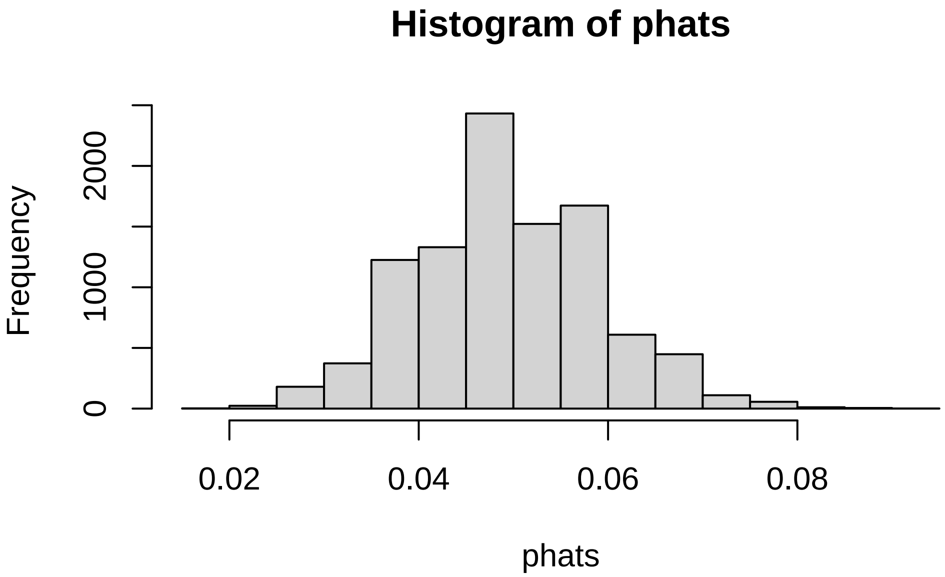
```r
prob <- pnorm(p + 0.01, p, SE) - pnorm(p - 0.01, p, SE)
```

The probability that the sample of 500 voters being within 0.01 of the true population proportion is 0.695.

c.

```r
set.seed(123)

samples <- 10000
X <- rbinom(samples, n, p)
phats <- X / n
hist(phats)
```



The mean of our distribution is 0.0499 and the standard deviation is 0.00965. This is very close to our calculated values, where we expected to get a mean of 0.05 and a standard error of 0.00975. These results support my answers from the previous question.

4.

```r
health <- source("https://www.openintro.org/data/R/healthcare_law_survey.R")$value
```

```r
approve <- table(health$response)["approve"]
table(health$response)
```

```
 approve disapprove      other 
     614         842         47 
```

614 people approved of this policy. The sample proportion is 0.409. This is the proportion of people that approved of this policy.

```r
CI <- prop.test(approve, nrow(health))
lower <- CI$conf.int[1]
upper <- CI$conf.int[2]
```

The 95% confidence interval is (0.384, 0.434).

We are 95% confident that the true proportion of people who approve of this policy is between 0.384 and 0.434. If we take a lot of samples of the same size and compute their confidence interval, 95% of those CIs will contain the true proportion.

5.

a. 765 adults in the United States is the population.
b. The parameter being estimated is the proportion of United States adults that cannot cover a $400 unexpected expense without borrowing money or going into debt.
c. The point estimate ($\hat{p}$) is $\frac{322}{765} \approx 0.421$.
d.

```r
CI95 <- prop.test(322, 765)
lower <- CI95$conf.int[1]
upper <- CI95$conf.int[2]
```

The 95% confidence interval is (0.386, 0.457).

e.

```r
CI90 <- prop.test(322, 765, conf.level = 0.9)
lower <- CI90$conf.int[1]
upper <- CI90$conf.int[2]
```

The 90% confidence interval is (0.391, 0.451).

Lowering the confidence interval makes the interval smaller (more narrow), but lowers the confidence that the true proportion is contained in the interval.

f.

```
phat <- 322/765
CId <- prop.test(phat * 1530, 1530)
lower <- CId$conf.int[1]
upper <- CId$conf.int[2]
```

The confidence interval is (0.396, 0.446). The interval becomes more narrow when we take more samples (larger n) from the population. It is more narrow compared to part (d).

6.

a. False. SE is not related to the proportion of American's that participated, but rather the absolute number of participants (n).
b. False, the n is in the denominator of the SE formula, so we should collect more data to decrease our standard error.
c. False. A lower confidence level means you get a smaller interval. A 100% confidence interval would include every number, which shows that if you increase the confidence level you get a wider interval.